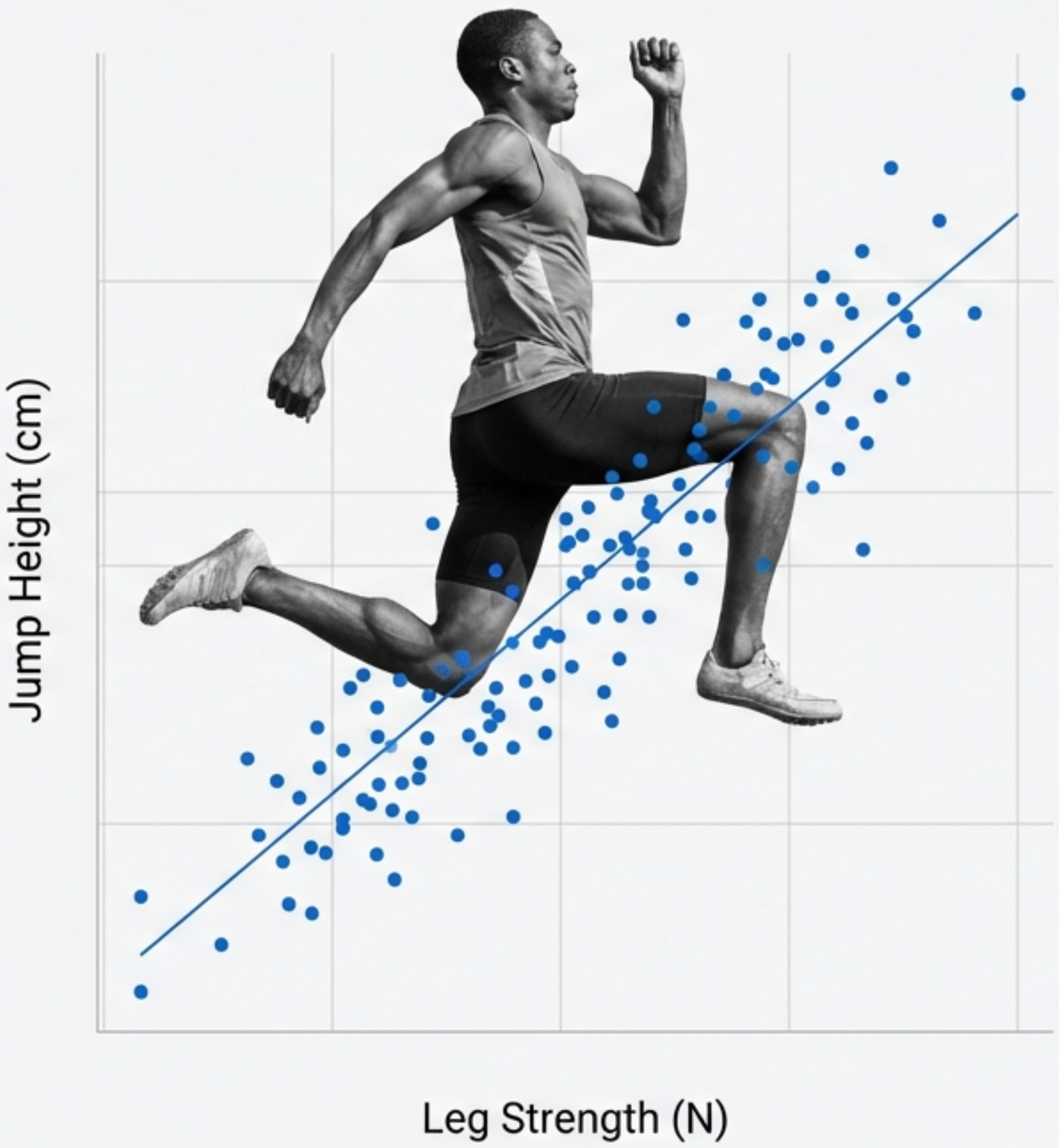


Correlation and Bivariate Regression in Movement Science

A case study on quantifying relationships and predicting performance

Many questions in Movement Science involve *relationships* between variables. *Does greater muscular strength predict higher vertical jump performance?*

In this study guide, we will answer that question by walking through the full statistical workflow: from visualizing raw data to building a *predictive mathematical model.*

NotebookLM

SCIENTIFIC WORKFLOW: CORRELATION AND BIVARIATE REGRESSION

1

First, we create a scatterplot to visualize the relationship and check for linearity.

→

2

Next, we compute the correlation coefficient (r) to quantify the strength and direction of the relationship.

→

3

We test for statistical significance to see if the result is distinguish-able from zero.

→

4

If prediction is the goal, we fit a regression model (R^2).

→

5

Finally, we check critical assumptions like linearity and homoscedas-ticity.

→

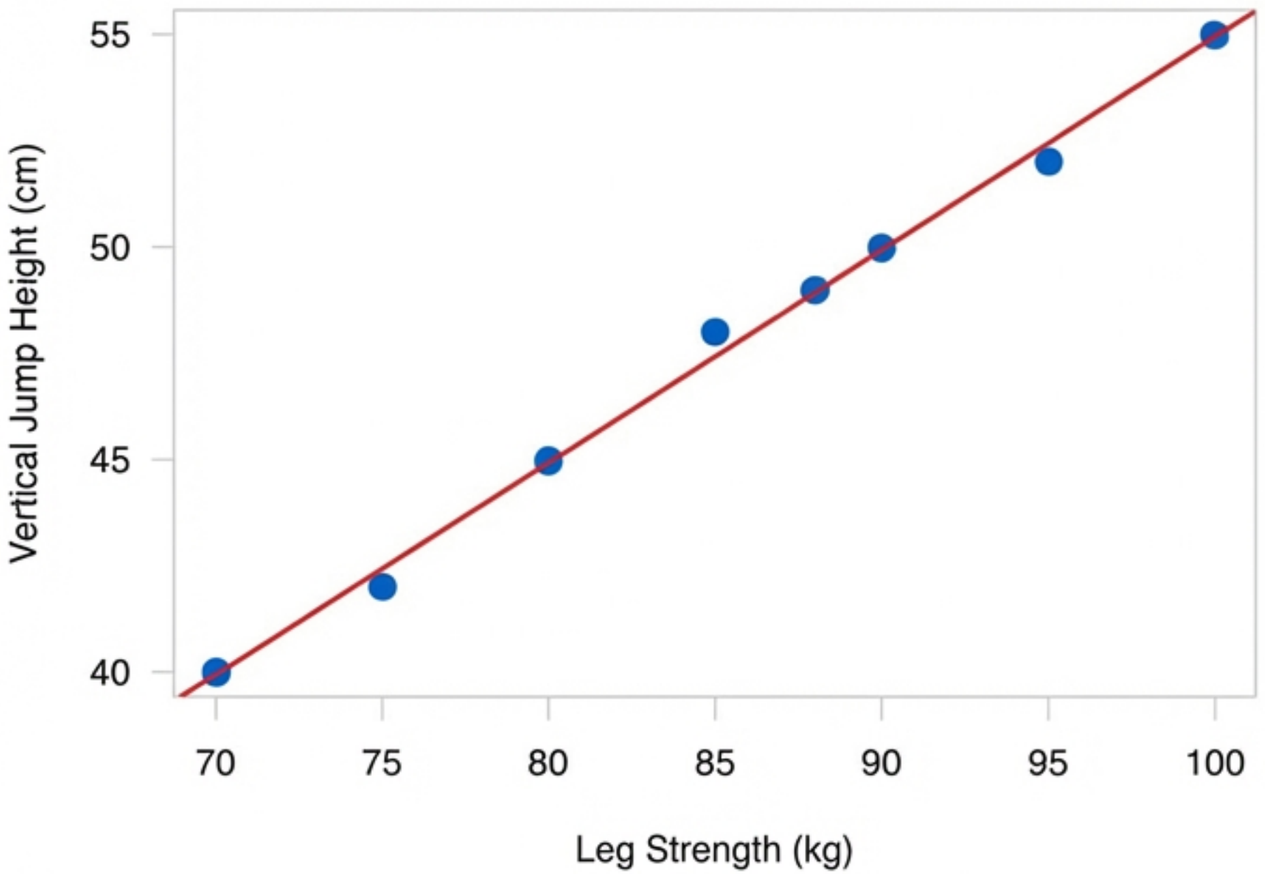
6

Interpret the results cautiously, remembering that correlation does not imply causation.

NotebookLM

Scientific Editorial

Phase 1: Visualizing the Data



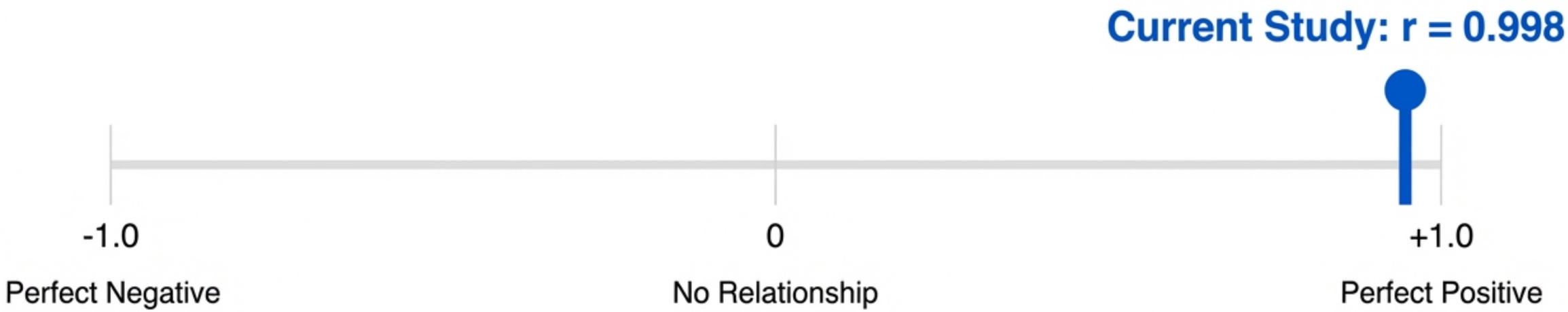
We begin by plotting our data: Leg Strength (kg) on the x-axis and Vertical Jump Height (cm) on the y-axis.

This step is essential because identical statistical values can arise from vastly different datasets—a phenomenon known as *Anscombe's Quartet*. Always visualize your data first.

The scatterplot reveals a clear *positive linear trend*, where points cluster tightly along an upward path.

NotebookLM

Quantifying the Relationship: Pearson's r



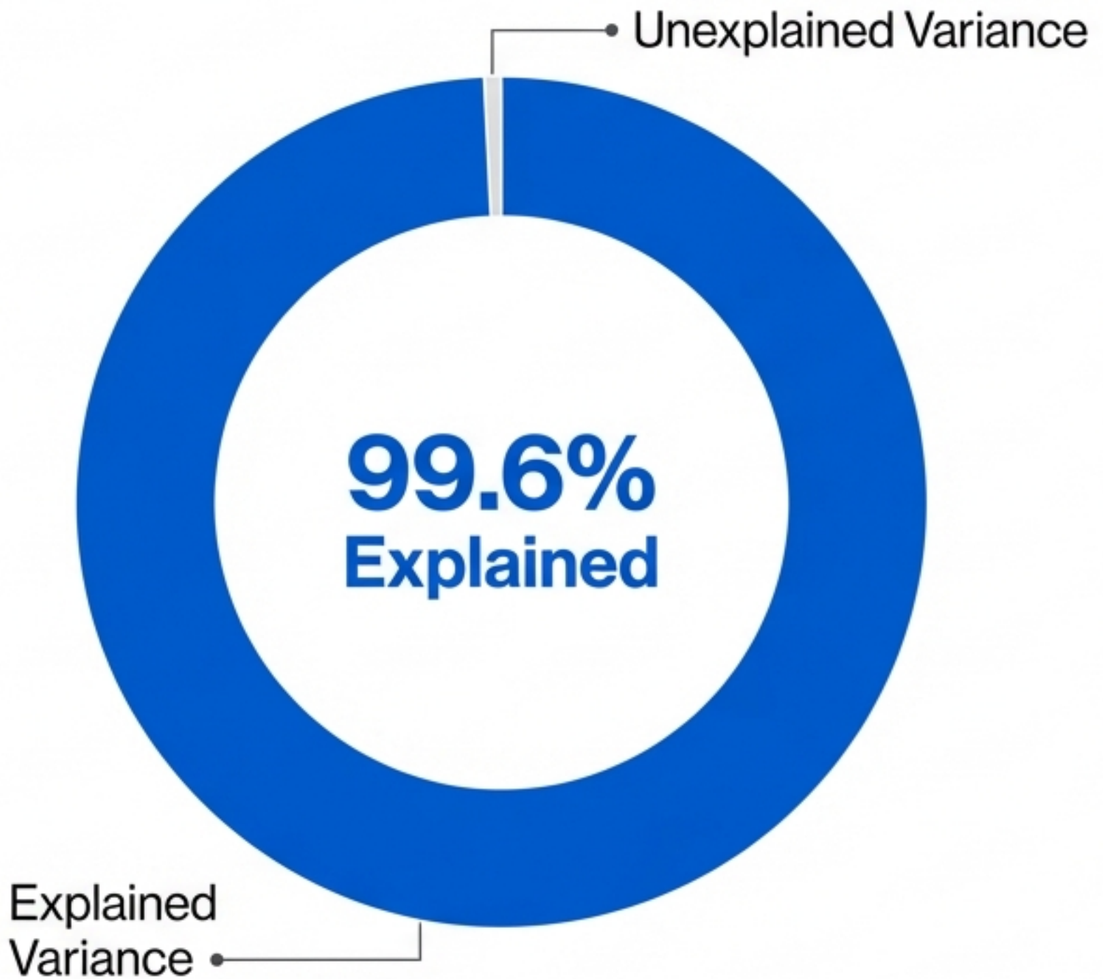
Pearson's correlation coefficient (r) measures the strength and direction of the linear relationship between two continuous variables.

For our athlete dataset, the computed r value is 0.998.

This indicates an *extremely strong, positive linear relationship*. As leg strength increases, vertical jump height tends to increase proportionally.

NotebookLM

The Coefficient of Determination (r^2)



The coefficient of determination (r^2) represents the proportion of variance in the outcome variable that is explained by the predictor.

By squaring our correlation (0.998^2), we get 0.996.

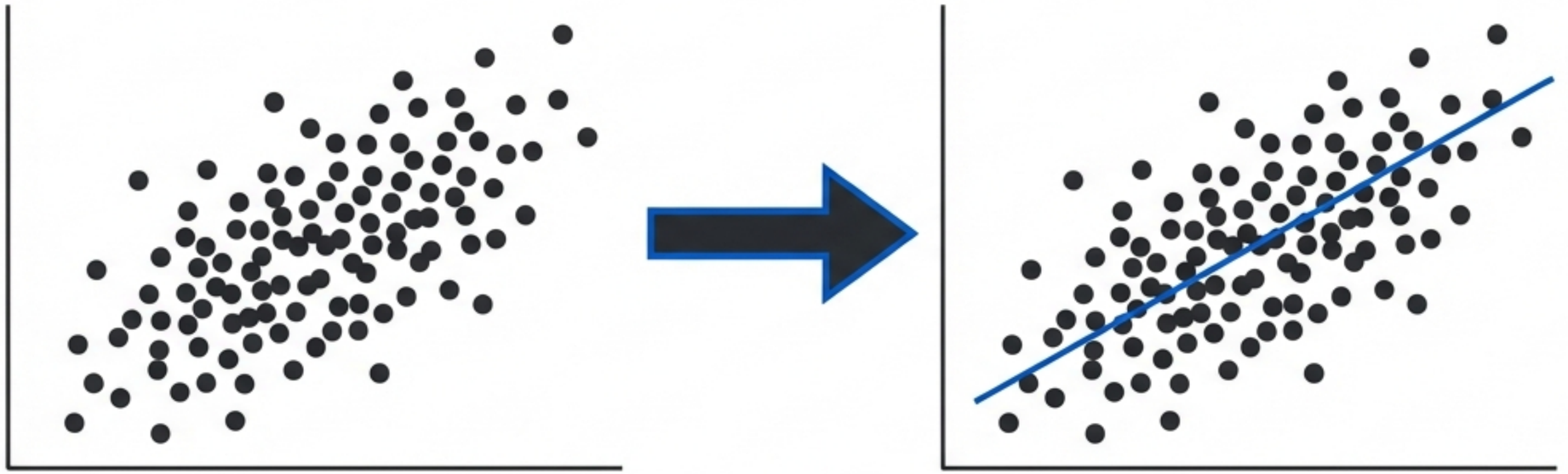
This means that 99.6% of the variability in vertical jump height is explained by differences in leg strength in this sample. The remaining 0.4% is due to measurement error or unmeasured factors like technique.

NotebookLM

Phase 2: From Correlation to Regression

Correlation: Describing the connection.

Regression: Modeling the prediction.



While correlation quantifies the strength of a relationship, regression models the relationship mathematically.

Bivariate linear regression fits a straight line to the data, producing an equation that allows us to predict an outcome (Dependent Variable Y) from a predictor (Independent Variable X).

NotebookLM

The Regression Equation

b (Slope): The slope represents the change in the outcome (Y) for every one-unit increase in the predictor (X).

$$\hat{y} = a + bx$$

a (Intercept): The intercept is the theoretical value of Y when X is zero.

This mathematical model ensures the line passes through the geometric center of the data $(\bar{x}, \bar{y})$.

NotebookLM

Interpreting the Slope: The Real-World Meaning

$$\hat{y} = 4.938 + 0.500x$$

In our athlete example, the slope is 0.500.

This means that for every single kilogram of leg strength an athlete adds, we expect their vertical jump height to increase by 0.5 centimeters on average.

The intercept (4.938 cm) is the predicted jump height if strength were zero. Since no athlete has zero strength, this value is theoretical and should not be over-interpreted.

NotebookLM

Making a Prediction

New Athlete
Strength: 92 kg

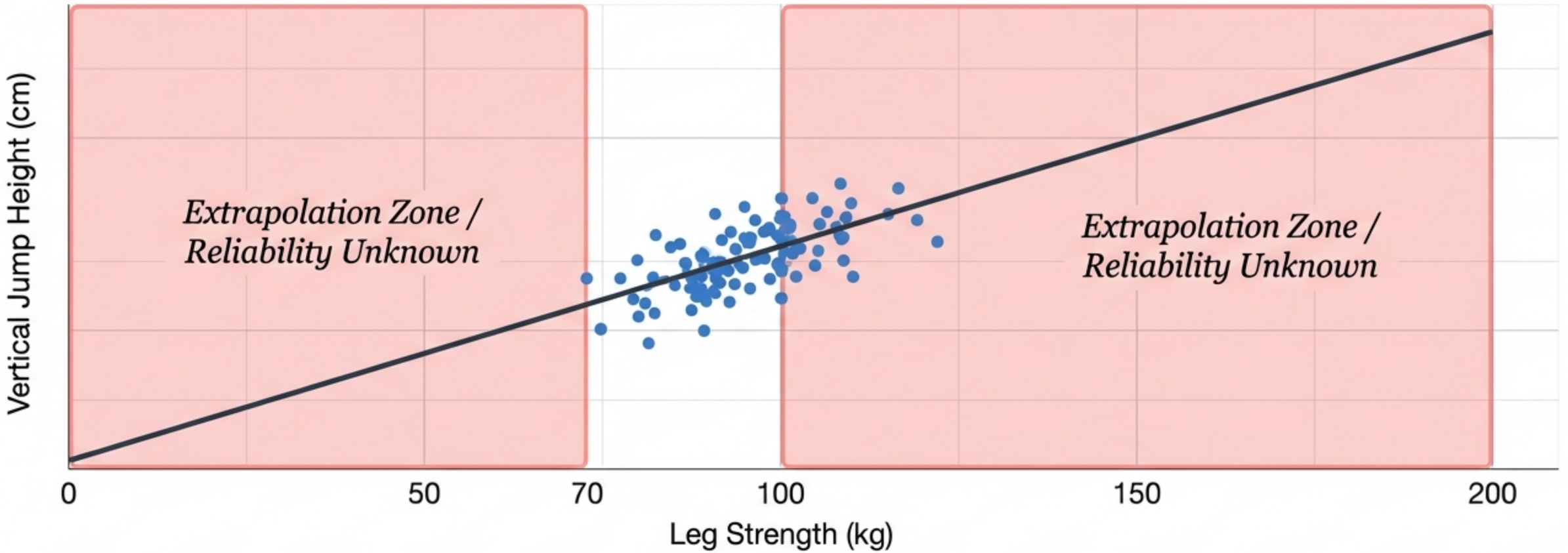
$$\hat{y} = 4.938 + 0.500(92) \longrightarrow$$

Predicted Jump: 51 cm

By plugging 92 kg into our regression equation, the model predicts a vertical jump height of approximately 51 cm. This demonstrates how regression allows us to estimate performance for values not originally in our dataset.

NotebookLM

The Danger of Extrapolation

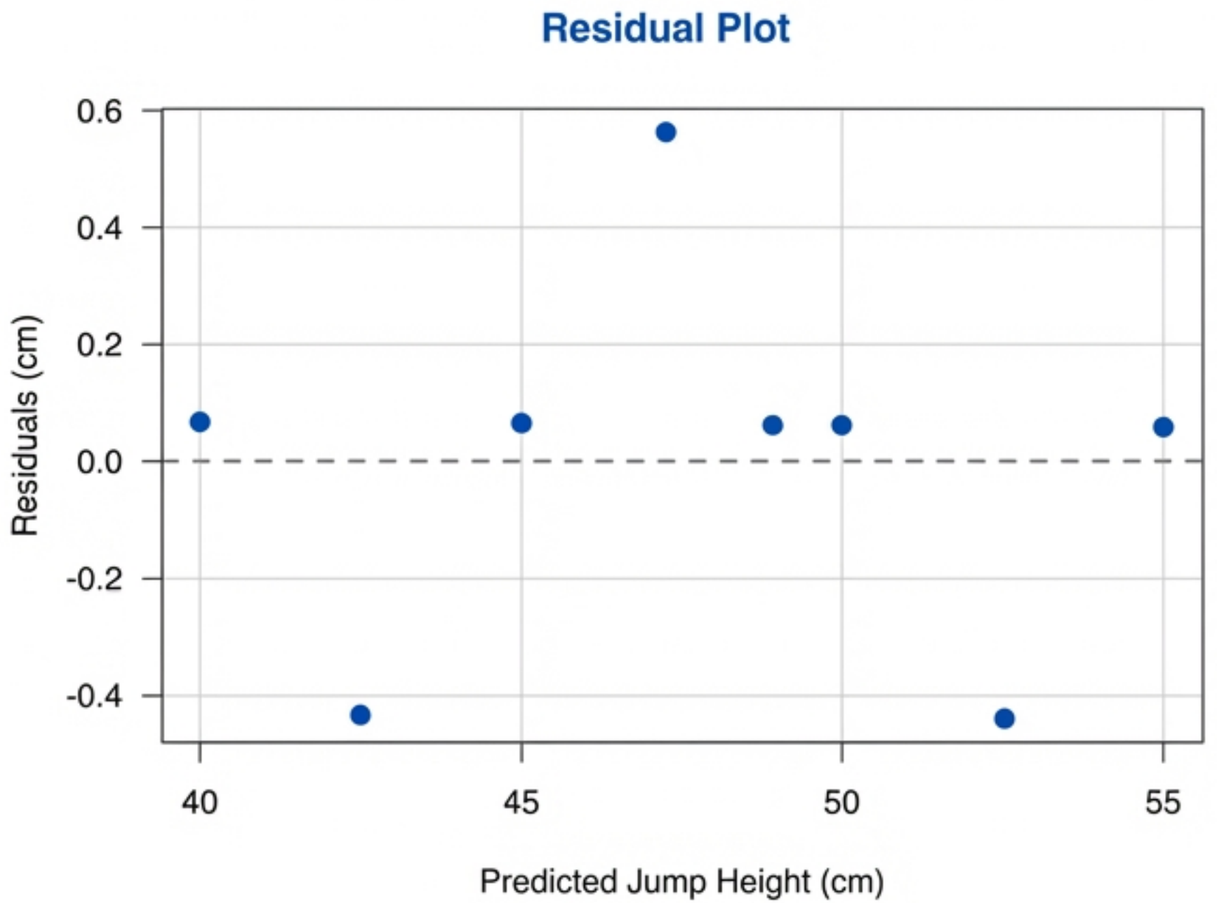


We must never use the regression equation to predict values outside the range of our observed data. Our data describes athletes with strength between 70 kg and 100 kg. Predicting for an athlete with 200 kg or 50 kg of strength is *extrapolation*—it is guessing, not science, as the relationship may change at those extremes.

NotebookLM

Phase 3: Validation with Residuals

A residual is the difference between what our model predicted and what actually happened (residual = y - ŷ).

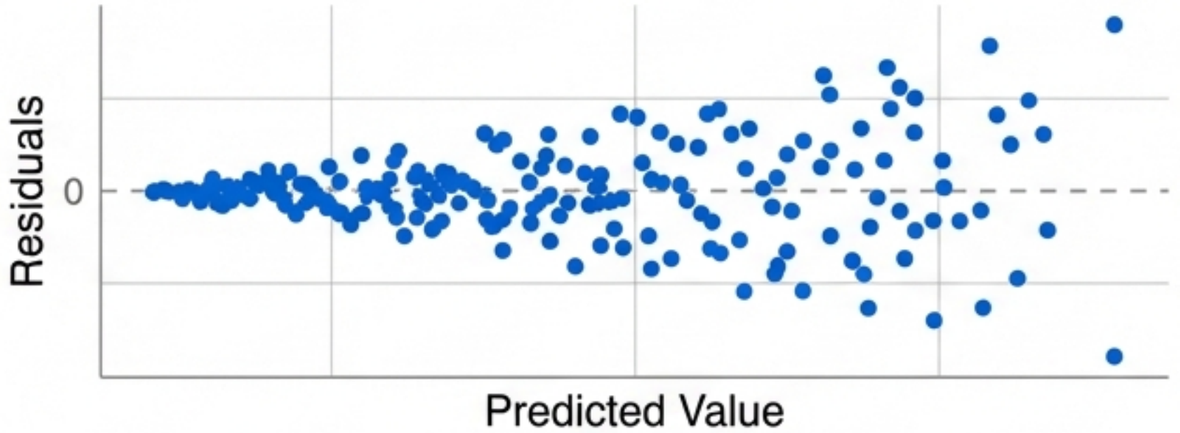


A good residual plot looks like random noise. The plot shown here displays a random scatter around zero with no obvious pattern, confirming that a linear model was appropriate for this data.

NotebookLM

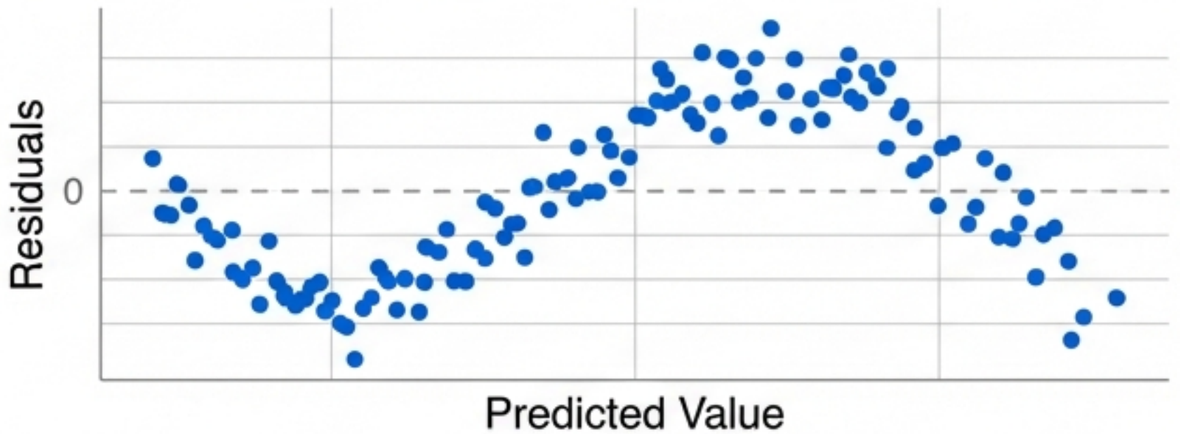
Critical Assumptions: Linearity & Homoscedasticity

Violation of Homoscedasticity



Homoscedasticity means the error remains constant. If the spread of errors gets wider as values increase (a funnel shape), the assumption is violated.

Violation of Linearity

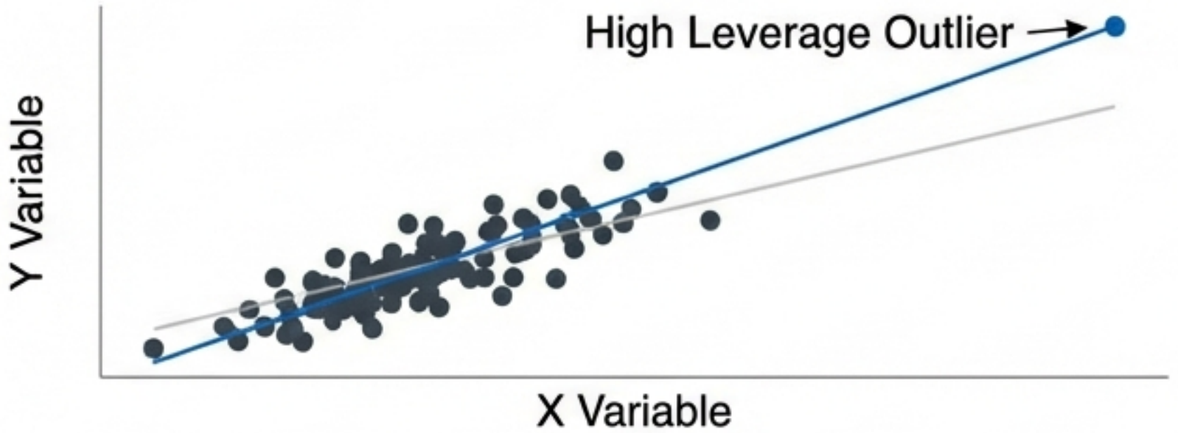


Linearity requires that the relationship is actually a straight line. If residuals show a curved pattern, a linear model is misleading.

NotebookLM

Critical Assumptions: Outliers & Independence

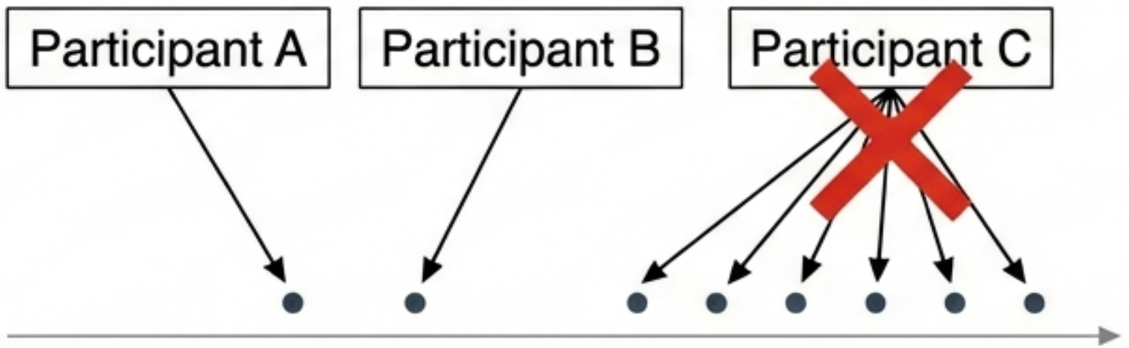
Outliers: High Leverage & Influence



A single athlete with extreme data can disproportionately influence the model. We must examine plots for high leverage points that distort the slope.

Independence of Observations

Observations must be independent. In movement science, this usually means each data point comes from a different participant, rather than measuring the same person multiple times.



NotebookLM

The Golden Rule: Correlation ≠ Causation

The Confounding Variable (Heat).

A strong statistical link does not prove one variable causes the other.

For example, ice cream sales and drowning deaths are positively correlated, but ice cream does not cause drowning. Both are influenced by a third, confounding variable: hot weather.

Similarly, body mass and VO_2max may correlate simply because larger people have larger hearts, not because mass itself causes higher oxygen uptake.

NotebookLM

Summary & Key Takeaways

1	We visualized the data to confirm a linear trend before calculating anything.
2	We quantified the relationship with Pearson's r (0.998) and r^2 (99.6%).
3	We built a regression model ($\hat{y} = a + bx$) to make predictions about athlete performance.
4	We validated the model by checking residuals for linearity and homoscedasticity.

Final Thought: *While these tools allow us to predict and model relationships, we must always remember that correlation is descriptive and does not strictly imply causation.*

NotebookLM